Predicting Cirrhosis Patient Survival Rate: Integrating SVM and KNN Algorithms for Enhanced Clinical Decision-Making ECS784P

***Abstract*—** Cirrhosis, a consequence of prolonged liver damage, presents significant challenges in clinical care, often necessitating liver transplant as a life-saving option. Survival analysis plays a crucial role in medical decision-making, particularly in predicting life expectancy for patients. This paper explores the application of Machine Learning models, including k-nearest neighbours (KNN), and support vector machines (SVM), in predicting liver transplant survival rates. Utilizing data sourced from a seminal Mayo Clinic study conducted from 1974 to 1984 on primary biliary cirrhosis (PBC), our study investigates the efficiency of machine learning algorithms in predicting survival outcomes for cirrhotic patients.

Keywords: Liver transplant, Survival analysis, K-nearest neighbours, Support vector machines, Personalized medicine.

I. Introduction

Chronic liver diseases, pose significant health challenges globally, contributing to substantial mortality rates. Understanding the factors influencing the survival rates of cirrhosis patients is important for effective medical management and resource allocation. In this study, we investigate the predictive modelling of survival rates among cirrhosis patients, leveraging machine learning techniques applied to a comprehensive dataset consisting of demographic, clinical, and laboratory features. The dataset used in our study originates from the Mayo Clinic experiment sourced from Kaggle.

The target variable in our analysis represents the status of patients, categorized into three classes: C (censored), CL (censored due to liver transplantation), and D (deceased).

We explore the application of two prominent machine learning algorithms: Support Vector Machines (SVM) and k-Nearest Neighbors (KNN). SVM is renowned for its ability to construct optimal decision boundaries in high-dimensional feature spaces, while KNN offers simplicity and adaptability to diverse datasets. Our objectives are as follows:

1. Develop and implement advanced machine learning algorithms tailored for multi-class classification to accurately predict cirrhosis survival rate.
2. To assess the predictive performance of different algorithms, providing insights into their strengths and limitations.
3. Employ advanced feature selection techniques to identify the most relevant features that maximize overall model performance while minimizing overfitting.

II. Literature Review

In the paper titled "The Albumin-Bilirubin Score as a Predictor of Liver-Related Mortality in Primary Biliary Cholangitis with Compensated Cirrhosis" the authors aimed to evaluate the prognostic performance of the albumin-bilirubin (ALBI) score and Mayo Risk Scores in PBC patients. They highlighted the importance of developing reliable prognostic tools that identify the degree of liver injury based on clinical and laboratory variables. These scores can be utilized to predict survival outcomes accurately [1].

The research paper named ‘Analysis and Prediction of Liver Cirrhosis Using Machine Learning Algorithms’ *by* Lalithesh D Sawant; Raghavendra Ritti [2] is being discussed. The report conducted a comparative study on prediction of Liver Diseases using several machine learning algorithms such as Support Vector Machines and K-Nearest Neighbours (KNN) with an accuracy of approx. 75%. This study emphasized on creating a subsets of data by distinguishing external and match-related features. This report suggests that by leveraging KNN and including additional variables, we can enhance our understanding of factors influencing survival outcomes [2].

In their paper titled "Diagnosis of Liver Disease using Machine Learning Models" A. Sivasangari and Baddigam Jaya Krishna Reddy present an innovative approach utilizing machine learning algorithms to predict and classify liver diseases. The study focuses on employing three distinct algorithms, namely Support Vector Machines (SVM), Decision Trees (DT), and Random Forest (RF), to analyze different attributes associated with liver diseases. The finding proves the potential of machine learning models, particularly SVM, in accurate diagnosis and classification of liver diseases [3].

The paper "A Novel Machine Learning Approach Using Boosting Algorithm for Liver Disease Classification" addresses key preprocessing steps for enhancing liver disease classification. It employs missing data imputation, label encoding, and redundant values removal to optimize the dataset. Additionally, the study tackles class imbalance using the Synthetic Minority Over Sampling Technique (SMOTE) for balanced sample representation, aiming

to improve predictive accuracy and mitigate bias in liver disease classification models [4].

The paper "Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients" investigates liver disease prediction. It employs various data cleaning techniques such as imputation, label encoding, and outlier elimination for improved performance. Genetic Algorithm with XGBoost is used to select optimal attributes, and different classification algorithms are applied for prediction. Performance metrics like accuracy, precision, recall, and f-measure are used to evaluate algorithm effectiveness [5].

III. Data Management

A. Cirrhosis Dataset

The dataset used in our study originates from the Mayo Clinic experiment conducted between 1974 and 1984, focusing on primary biliary cirrhosis (PBC) of the liver. The shape of the dataset is (7905, 20) and contains demographic information, clinical features, and laboratory measurements. We obtained the dataset from Kaggle.

Variable Name	Description	Units
ID	Unique identifier	None
N_Days	Number of days between registration and analysis	Days
Status	Status of the patient	None
Drug	Type of drug	None
Age	Age of the patient	Days
Sex	Gender of the patient	None
Ascites	Presence of ascites	None
Hepatomegaly	Presence of hepatomegaly	None
Spiders	Presence of spider angiomas	None
Edema	Presence of edema	None
Bilirubin	Serum bilirubin level	mg/dl
Cholesterol	Serum cholesterol level	mg/dl
Albumin	Albumin level	gm/dl
Copper	Urine copper level	ug/day
Alk_Phos	Alkaline phosphatase level	U/liter
SGOT	SGOT level	U/ml
Tryglicerides	Triglycerides level	None
Platelets	Platelets per cubic	ml/1000
Prothrombin	Prothrombin time	s
Stage	Histologic stage of disease	None

Figure 1. Dataset Variable Table

The target 'Status', denoting the patient's status categorized as censored (C), censored due to liver transplant (CL), or deceased (D).

Figure 2. Dataset

B. Data pre-processing

The implemented preprocessing pipeline contains handling missing values, converting categorical variables, and removing redundant features to ensure the quality and integrity of the dataset. Missing values in the 'Cholesterol', 'Copper', 'Alk_Phos', 'SGOT', 'Prothrombin' and 'Triglycerides' columns, denoted by 'NaN', were imputed with their mean or median values. For 'Cholesterol', the mean value (350) was used for imputation, while the median value (104) was utilized for 'Tryglicerides'. For the 'Stage', 'Copper', 'Alk_Phos', and 'SGOT' columns, missing values were imputed with the mode, representing the most frequent value within each respective column. The mode values were determined to be 2, 83 ug/day, 1187 U/litter, and 114 U/ml, respectively. Categorical variables such as 'Drug', 'Ascites', 'Hepatomegaly', and 'Spiders' had missing values replaced with the mode, representing the most frequently occurring category within each column. For instance, 'Drug' was imputed with 'D-penicillamine', while 'Ascites', 'Hepatomegaly', and 'Spiders' were imputed with 'N' (No).

The 'id' column, containing unique identifiers, was removed as unique identifiers do not contribute to model learning and are therefore deemed unnecessary for analysis. To ensure data integrity and avoid redundancy, we conducted a check for duplicate entries in the dataframe. This was achieved using the df.duplicated().sum() function, which provides the total count of duplicate records. After inspection, it was confirmed that there are no duplicate entries present in the dataset.

C. Exploratory Data Analysis

Analysing the target distribution in a dataset is vital for detecting class imbalances, guiding the selection of appropriate features. Upon plotting the distribution of the three target classes, it is observed that the majority of cases fall under the category 'C' (censored) with 4965 occurrences, followed by 'D' (deceased) with 2665 instances, and 'CL' (censored due to liver transplantation) with 275 occurrences, stating the need to address class imbalance for accurate predictions.

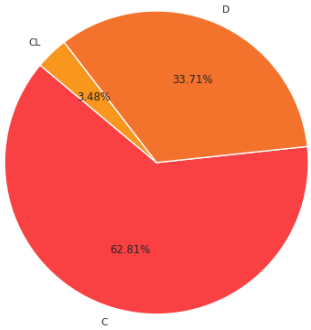

Figure 3. Target Distribution

Distribution of Numerical Features: Univariate analysis aids in understanding the distribution and

characteristics of individual variables within a dataset, giving us insights about the spread, skewness and central tendency.

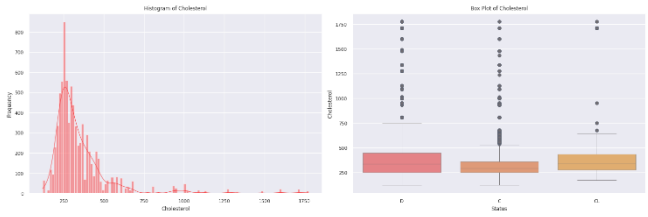

Figure 4. Distribution of Cholesterol

After examining the box plot distribution of the 'Cholesterol' feature across the three classes (D, CL, and C), it's evident that the majority of values fall within the range of 250-500 for all classes. Specifically, the mean values for 'Cholesterol' are approximately 260, 220, and 250 for classes D, CL, and C respectively. The histogram of 'Cholesterol' is skewed to the left, suggesting that the majority of patients have lower cholesterol levels. We can also see the outliers denoted by grey dots.

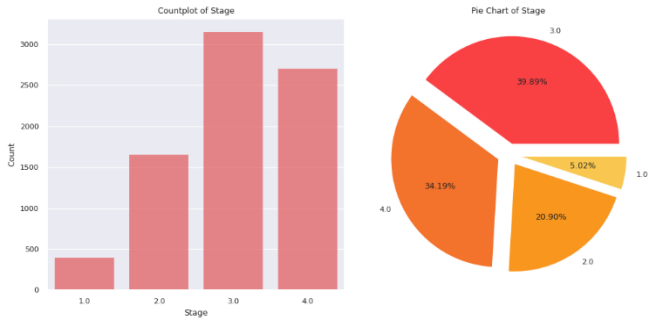

Figure 5. Distribution of Histologic Stage

The distribution of the histologic stages (1, 2, 3, 4) in the above charts reveals that the majority of data values belong to Stage 3, followed by Stage 4 and Stage 1 has the least data points. Approximately 3200 instances, accounting for 39.89% of the dataset, fall under Stage 3, while approximately 2600 instances, representing 34.19% of the dataset, belong to Stage 4. This indicates a higher prevalence of patients in advanced stages (Stage 3 and Stage 4) compared to earlier stages (Stage 1 and Stage 2).

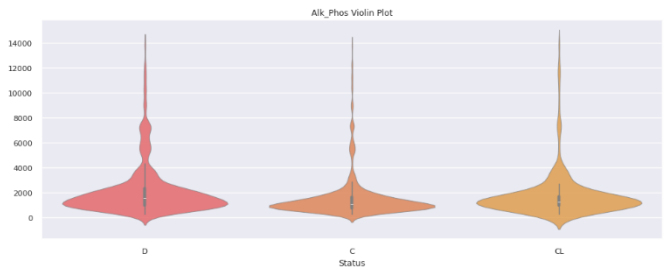

Figure 6. Violin Plot of alkaline phosphatase

This distribution of target classes is assessed in terms of alkaline phosphatase with the help of violin plot. We can observe that majority of the data points for Class D and CL are in the range of 0- 4000, whereas for class C the range is between 0-2000. The thin tail indicates that there are fewer observations with higher values compared to the majority of the data. This indicate skewness towards lower values with a few outliers towards higher values.

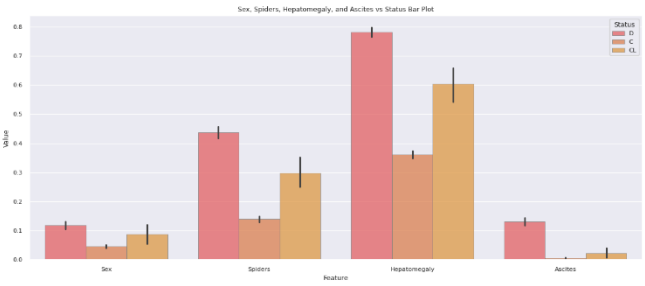

Figure 7. Distribution of categorical features ('Sex', 'Spiders', 'Hepatomegaly', 'Ascites') across different classes of the target variable

Multivariate analysis: By visualizing the distribution of categorical features ('Sex', 'Spiders', 'Hepatomegaly', 'Ascites') across different classes of the target variable, we can identify potential associations between these variables and the survival status of cirrhosis patients. Across all four variables, class D consistently exhibits the highest number of instances and class C has the least instances. This suggests an imbalance in class distribution, with class C being underrepresented relative to the other classes.

Based on the research conducted by Hung Leng Kaan, Khek Yu Ho [1],we have included two features based on prognostic scores, the albumin-bilirubin (ALBI) score and the Mayo risk score, to evaluate the risk of mortality. The ALBI score, is based on serum albumin and total bilirubin levels, while the Mayo risk score, developed by the Mayo Clinic, considers age, total bilirubin levels (TBIL), prothrombin time (PT), presence of edema, and albumin levels (ALB). The ALBI_score is classified into 3 main classes as suggested in the referred paper.

D. Outlier Detection and Removal

We have computed Z-Scores for numerical feature in the dataset using the formula: $Z = \sigma|X-\mu|$, where X is the data point, μ is the mean, and σ is the standard deviation. This normalization process standardizes the data and allows for the detection of outliers based on their deviation from the mean. We have a common threshold of Z-Score greater than 6 is considered as an outlier.

E. Feature Selection

We created a correlation matrix to examine the relationships between numeric features in our dataset.

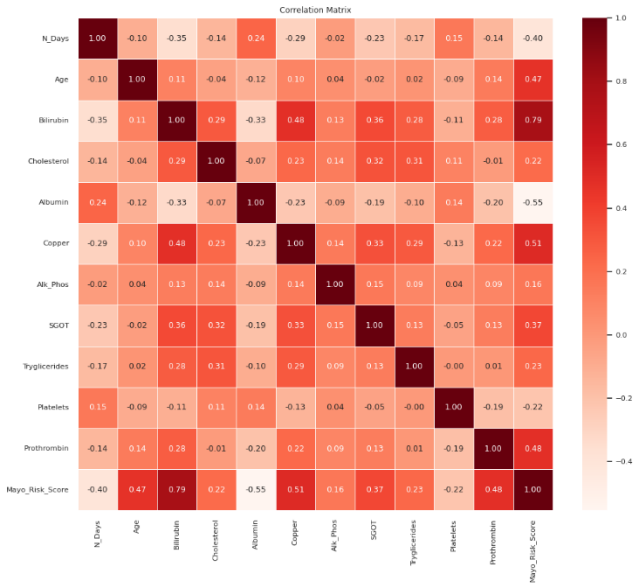

Figure 8. Correlation matrix of the entire dataset

The darkest colour indicates that the two features have a high correlation, whereas the lightest colour (White) indicates that the features have a low correlation score. We observed a significant correlation between 'Mayo_Risk_Score' and 'Billublin' (0.79) and lower corelation between 'Albumin', and 'Mayo_Risk_Score' (-0.55).

Furthermore, we have implemented binary encoding and one-hot encoding techniques to transform categorical variables into numerical representations. In 'Ascites', 'Hepatomegaly', and 'Spiders' with binary encoding 'Y' was assigned value (1), while 'N' was labelled as 0. One-hot encoding was used for categorical variables such as 'Drug', 'Edema', and 'Stage'. This expanded each categorical variable into different binary columns having an unique value. For instance, 'Drug' was expanded into ('Drug_D-penicillamine' and 'Drug_Placebo'), 'Stage' into four columns ('Stage_1.0', 'Stage_2.0', 'Stage_3.0', and 'Stage_4.0').

F. Data Normalization

We normalized the dataset using Min-Max scaling, this transformed the data so that it falls within a specific range. We did this to ensure that all features contribute equally. This helps to avoid large scale features from dominating the learning process. We compute the difference between the maximum and minimum values.

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Where, X is the original value of the feature. $X_{\min}$ is the minimum value of the feature in the dataset, $X_{\max}$ is the maximum value of the feature in the dataset, X_{scaled} is the normalized value of the feature, which falls within the range.

In our analysis, we employed Min-Max scaling to promote fairness in feature contributions and to mitigate potential biases introduced by variations in feature scales. Additionally, Min-Max scaling enhances the interpretability of the model's coefficients and makes more meaningful comparisons between features.

G. Handling Class Imbalance

The distribution of the target classes was uneven, with class CL comprising only 275 instances being underrepresented compared C(4965) and D(2665). This could create biased model performance. To avoid this we implemented oversampling technique, known as Synthetic Minority Oversampling Technique, which generates synthetic samples along the line segments joining neighbouring minority class instances. Applying SMOTE resulted in an equal representation of each class with 4942 instance.

H. Feature Importance

Feature importance gives us insights about which features contribute the most to the model's predictive performance. Random Forest is a powerful ensemble learning technique that combines multiple decision trees to make predictions. It calculates feature importance based on the decrease in impurity (e.g., Gini impurity or entropy) that results from splitting nodes, which helps in understanding the relationships between features and the target variable.

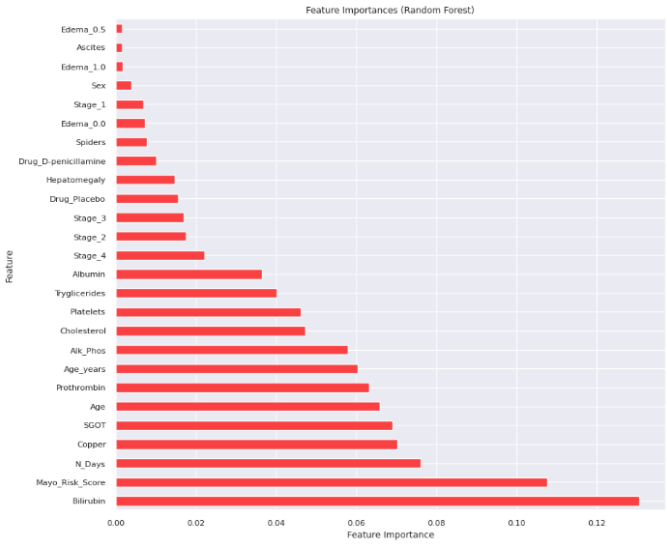

Figure 9. Feature Importance using Random Forest Classifier

The features 'Ascites', 'Spiders' and 'Sex' are with the least significance in influencing the outcome. Therefore, these features were dropped from the dataset.

A. Hypothesis Testing

Hypothesis testing is a statistical method used to make inferences about a feature based on the data. For the first hypothesis testing we investigate whether there is a difference in survival rates between patients treated with D-penicillamine and those treated with a placebo.

(H0): There is no difference in survival rates between patients treated with D-penicillamine and those treated with a placebo.

(H1): There is difference in survival rates between patients treated with D-penicillamine and those treated with a placebo.

Chi-square statistic: 0.024271754428908148
p-value: 0.9879374655506878

We used the chi-square test for independence, which assesses whether there is a significant association between two categorical variables. The results of our hypothesis test revealed a chi-square statistic of approximately 0.0243 and a p-value of approximately 0.9879, with 2 degrees of freedom. We failed to reject the null hypothesis.

IV. METHODOLOGY

We have utilized two machine learning algorithms, to predict the survival of patients with cirrhosis. Support Vector Machines (SVM) and k-Nearest Neighbours (KNN) are both supervised machine learning algorithms that can be effectively used for multi-class classification.

A. Support Vector Machines (SVM)

SVM is a powerful supervised learning algorithm that constructs a hyperplane in a high-dimensional space to separate data points into different classes. It works by finding the optimal hyperplane that maximizes the margin between classes, thereby achieving good generalization performance. The decision function of SVM can be represented as:

$$f(x) = \text{sign}(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b)$$

Where, $f(x)$ represents the predicted class label, N is the number of support vectors, α_i are the coefficients, y_i are the class labels, x_i are the support vectors, x is the input feature vector, K is the kernel function, and b is the bias term. The input feature vector x represents a single data point, consisting of the values of the features used for prediction. When we are predicting the status of a patient, x would include the values of features for that patient. The class label y would indicate the target variable, Status. It could be one of the following categories: censored (C), censored due to liver transplant (CL), or deceased (D).

We selected Support Vector Machines (SVM) because it provides offer versatility in handling both linear and non-linear classification tasks, making them well-suited for capturing complex relationships in liver disease prediction. As our dataset is medium-sized, SVM is lesser prone to overfitting, especially with small or mid-sized dataset. It also provides flexible parameter tunning with which we can fine tine our predictive accuracy. Also, SVM has been successfully applied in several medical domains, including disease diagnosis and treatment prediction. In general, the versatility, robustness to overfitting, margin maximization properties, tuning flexibility, and proven track record in medical applications makes it an ideal candidate for our dataset..

B. KNN

KNN is a supervised learning algorithm used for classification tasks. The algorithm works by finding the k nearest data points in the training set to a given query point and assigning the most common class label among those neighbours. It will calculate the distance between the new data point and all other data points in the training set using a distance metric (e.g., Euclidean distance, Manhattan distance). Select the k-nearest data points based on the calculated distances. And assign the majority class among the k-nearest neighbours as the predicted class for the new data point. The equation to calculate the distance between two data points X and Y using Manhattan distance is given by:

$$\sum_{i=1}^{n} |X_i - Y_i|$$

Where, X_i and Y_i are the i^{th} features of data points X and Y respectively. n is the number of features. In our Dataset, X_i represents a data point (a patient with cirrhosis) in the feature space, and Y_i represents the corresponding class label (survival status). X represents the matrix contains the features (such as age, sex, presence of ascites, hepatomegaly, etc.) of the patients, and the Y contains the corresponding class labels (CL, C, or D).

In the case of KNN, we included several values of k, several distance metrics such as Euclidean distance/ Manhattan distance. Adjusting hyperparameters such as the number of neighbours (n_neighbors), the distance metric (metric), and the weighting scheme (weights), we can fine-tune the KNN model to achieve better predictive accuracy.

We have performed hyper parameter tuning using GridSearchCV technique. The process involved searching through a predefined grid of hyperparameter values and selecting the combination that maximizes the performance of the model. In our case, the best parameters for the KNN algorithm were {'metric': 'manhattan', 'n_neighbors': 1, 'weights': 'uniform'}. This suggests that the algorithm achieved maximum accuracy (approximately 89.01%) when considering only the closest neighbour and using the Manhattan distance metric to measure similarity between data points. The uniform weighting scheme depected that all neighbours have an equal vote in the prediction process. As our medical dataset may contains localized patterns, K-Nearest Neighbours (KNN) performs well in capturing localized patterns and clusters within the dataset. Our dataset also contained several outliers, KNN is robust to outliers because it relies on local similarity measures rather than assuming global data distribution. In general, their simplicity, non-parametric nature, and robustness to irregular data distributions are particularly advantageous in healthcare settings.

V. TESTING AND RESULTS

A. ROC curve

The ROC curve is a graphical representation of the performance of a binary classifier system and it illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) as the discrimination threshold of the classifier is varied. The area under the ROC curve (AUC-ROC) values indicate the classifier's performance in distinguishing between the three classes

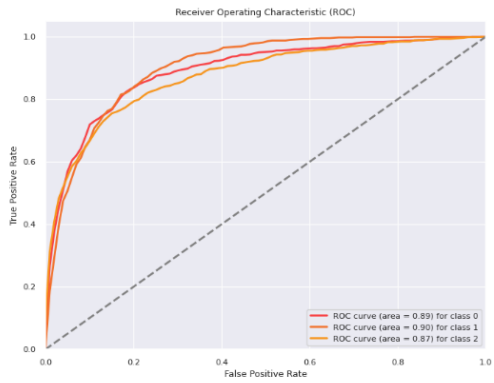

Figure 10. Receiver Operating Characteristics (ROC) SVM

SVM ROC Curve Interpretation: AUC values of 0.89, 0.90, and 0.87 for classes 0, 1, and 2, respectively, suggest strong discriminatory power for each class. Class 1 exhibits the highest AUC value (0.90), indicating that the SVM classifier performs exceptionally well in distinguishing class 1 from the other classes.

KNN ROC Curve Interpretation: KNN also exhibits promising performance, with AUC values ranging from 0.87 to 0.97 across the three classes. The highest AUC of 0.97 for class 1, indicating its superior ability to classify instances belonging to class 1.

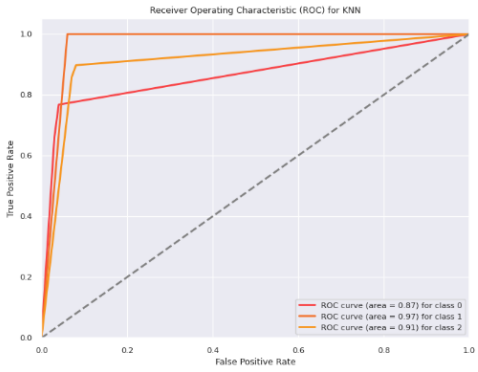

Figure 11. Receiver Operating Characteristics (ROC) KNN

The AUC values of 0.91 for class 2 and 0.87 for class 0 show KNN's effectiveness in discriminating between the class 2 and 0. Comparing the AUC values of SVM and KNN, we observe that both classifiers perform competitively.

Both SVM and, KNN excels particularly in classifying instances of class 1. The choice between SVM and KNN may depend on various factors such as computational efficiency, interpretability, and the specifications of the application.

B. Confusion Matrix

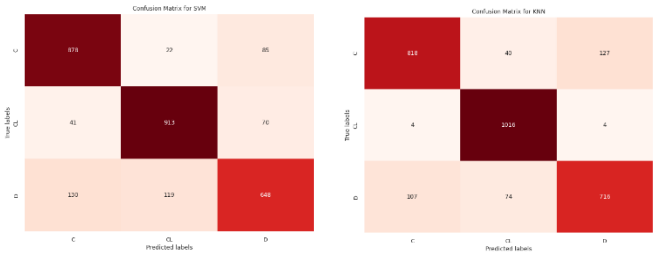

Figure 12. Confusion Matrix for SVM and KNN

SVM correctly predicts C 878 times, CL 913 times, and D 648 times. It misclassifies C as CL 72 times and as D 132 times. SVM misclassifies class C as class CL 41 times and as D 130, while the KNN model misclassifies class C as class CL 40 times and as D 127. KNN correctly predicts C 818 times, CL 1013 times, and D 710 times.

Both SVM and KNN demonstrate reasonably good predictive performance, with high accuracy in

predicting the classes (C, CL, and D). Both, SVM and KNN achieve higher accuracy in predicting class CL. KNN has very few instances of misclassification overall, indicating its robustness in this particular classification task.

C. Accuracy

For SVM, we utilized grid search with cross-validation to optimize the parameters, including the regularization parameter (C), kernel type, and gamma value. The best parameters obtained were {'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}. On the other hand, for KNN, we employed grid search to tune the number of neighbors (k), distance metric, and weighting scheme. The optimal KNN parameters were {'metric': 'manhattan', 'n_neighbors': 1, 'weights': 'uniform'}.

The KNN model outperformed the SVM model in terms of accuracy, achieving a higher accuracy of 89.01% compared to 83.93% of SVM.

The observed difference in accuracy between SVM and KNN highlights the importance of selecting the appropriate classification algorithm based on the characteristics of the dataset and the specific requirements of the task. While SVM is known for its ability to handle complex decision boundaries and high-dimensional data, KNN excels in scenarios where local information is crucial for classification.

VI. Concluding Remarks, Limitations and Future Scope

Liver disease progression is a critical healthcare concern globally, necessitating accurate predictive models for effective patient management. In this study, we analysed a comprehensive dataset encompassing various demographic, clinical, and histologic features of liver disease patients. After data preprocessing and feature engineering, we explored the distribution of target variables and performed extensive exploratory data analysis to identify key predictors of disease progression. Subsequently, we applied a range of machine learning algorithms, including, support vector machines, and KNN, to develop predictive models. Model performance was evaluated using appropriate metrics such as accuracy, Confusion Matrix and (AUC-ROC). Additionally, we addressed class imbalance issues through techniques such as oversampling. Furthermore, our models achieved promising accuracy rates, exceeding 89, underscoring their reliability in predicting liver transplant outcomes.

Nevertheless, there were certain limitations in the study. Firstly, the features in the dataset cannot fully capture the factors influencing the survival rate. Further research is needed to explore additional relevant features such as normal_sgot_range (serum glutamic oxaloacetic transaminase), normal_albumin_range, and disease severity score to improve model performance. Moreover, Combining our cirrhosis dataset with additional data sources such as genomic data, lifestyle factors, clinical imaging, electronic health records, population health data, and can refine our feature space, provide detailed insights, and potentially enhance the predictive performance and generalizability of our models for liver disease progression. Also, even though machine learning algorithms like SVM and KNN offer high predictive accuracy, their interpretability may be limited. Future research should focus on exploring the use of decision trees or rule-based models.

Future research to expand the scope of our project can focus on two key areas: Clinical Decision Support Systems (CDSS) and Patient Classification for Personalized Medicine. CDSS based on predictive modelling can provide real-time assistance to healthcare workers in patient management, and treatment planning. These systems could provide personalized insights about the progression of the disease, recommend test, and treatments based on the patients characteristics. Additionally, using predictive models for patient classification can help categorize patients into risk groups and treatment strategies can be customized accordingly. This could revolutionize healthcare department.

References

[1] Hung Leng Kaan, Khek Yu Ho; Diagnostic Challenges in Endoscopy. Dig Dis 11 December 2023; 41 (6): 833–834. https://doi.org/10.1159/

[2] L. D. Sawant, R. Ritti, H. N, A. Kodipalli, T. Rao and R. B R, "Analysis and Prediction of Liver Cirrhosis Using Machine Learning Algorithms," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023,

[3] A. Sivasangari, B. J. Krishna Reddy, A. Kiran and P. Ajitha, "Diagnosis of Liver Disease using Machine Learning Models," *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Palladam, India, 2020,

[4] N. Afreen, R. Patel, M. Ahmed and M. Sameer, "A Novel Machine Learning Approach Using Boosting Algorithm for Liver Disease Classification," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021,

[5] M. A. Kuzhippallil, C. Joseph and A. Kannan, "Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020